Historical Data Preservation and Interpretation Pipeline for Irish Civil Registration Records

Oya Beyan[1], PJ Mealy[1], Dolores Grant[2], Rebecca Grant[2], Natalie Harrower[2], Ciara Breathnach[3], Sandra Collins[2] and Stefan Decker[1]

[1] Insight @ NUIG, National University of Ireland Galway, Galway, Ireland
{ oya.beyan, pj.mealy, stefan.decker }@insight-centre.org
[2] Digital Repository of Ireland, Royal Irish Academy, Dublin, Ireland
{ d.grant, r.grant, n.horrower, s.collins }@ria.ie
[3] Department of History, University of Limerick, Limerick, Ireland
Ciara.Breathnach@ul.ie

Abstract. Semantic Web technologies give us the opportunity to understand today's data-rich society and provide novel means to explore our past. Civil registration records such as birth, death, and marriage registers contain a vast amount of implicit information, which can be revealed by structuring, linking and combining that information with other datasets and bodies of knowledge. In the Irish Record Linkage (IRL) Project 1864-1913, we have developed a data preservation and interpretation pipeline supported by a dedicated semantic architecture. This three-layered pipeline is designed to capture separate concerns from the perspective of multiple disciplines such as archivistics, history and data science. In this study, our aim is to demonstrate best practices in digital archives, while facilitating innovative new methodologies in historical research. The designed pipeline is executed with a dataset of 4090 registered Irish death entries from selected areas of south Dublin City.

Keywords: Knowledge Transformation Pipelines . Civil Registration Records . Linked Data. . Digital Archives .

1 Introduction

Semantic Web technologies give us the opportunity to understand today's data-rich society and provide novel means to explore our past. Civil registration records such as birth, death, and marriage registers contain a vast amount of implicit information about society's past, which can be revealed by structuring, linking and combining that information with other datasets and bodies of knowledge. In the Irish Record Linkage 1864-1913 (IRL) project[1], we adopt Semantic Web and Linked Data technologies to create a platform for storing and linking RDF descriptions of birth, death and marriage (BDM) records for Dublin (1864-1913) [1]. The aim of IRL project is

[1] http://irishrecordlinkage.wordpress.com

to create a knowledge base, which can serve to answer questions about the accuracy of officially reported maternal mortality and infant mortality rates.

Semantic web and linked data technologies encapsulate the explicit representation of meta-information accompanied by domain theories such as ontologies, which will enable the web to provide a qualitatively new level of services [2]. These technologies have various advantages for capturing and interpreting the civil registration records. RDF metadata enables one to generate different models of data representation for separate concerns or interpretations. Because the linked data is self-describing and explicitly defined in a machine-readable way, it can be linked to external data sets and infer potential relevancies.

Motivation and Related Work

Our motivation is to develop novel methods to explore and interpret historical data sets with semantic web technologies and Linked Data. Digital repositories provide a central access point and interactive tools for historical and contemporary data [3]. These repositories may serve diverse interest groups such as archivists, historians, journalists, public researchers and scholars. The developed knowledge infrastructure should satisfy different, sometimes conflicting perspectives and concerns, as well as support privacy of the data subjects.

In digital preservation semantic technologies are applied for dynamically discovering and invoking the most appropriate preservation services [4]. XML data standards considered as opportunity in addressing the digital preservation problems [5]. In Neptuno system semantic web technologies are applied to create a knowledge base for digital newspaper archives. Archive materials are described using the developed ontologies and semantic search module implemented to use conceptual elements to match information needs against archive contents [6].

In this study, we have developed a three-layer pipeline for storing, exploring and interpreting these Irish civil registration records. We demonstrate our concept with the infant deaths as the use case. Infant mortality is an important indicator of human welfare, national wealth and social conditions such as poverty and single motherhood [7,8]. The pipeline will initially include 444 death register pages, which equates to 4090 death entries recorded in two Registrar Districts to the South of Dublin City from the years 1870 and 1890.

2 Methods

In this work we applied semantic web technologies to digital archival domain. We followed the linked data principles and express the semantic of data with the developed ontologies.

Semantic Web Technologies offers us to a new approach to managing information and process, the fundamental principle of which is the creation and use of semantic metadata [9]. Linked data refers to set of best practices for publishing and connecting structured data on the Web. It creates links between diverse data sources and enables

to publish data in a machine readable way where its meaning is explicitly defined and linked to the other data sets [10]. A computer-based, shared, agreed formal conceptualisation is known as an ontology [11]. Ontologies are keystone technologies for meaningful and efficient interoperation of information systems. They involve shared concepts and represents externalization of semantics outside of the systems [12].

3 Data Preservation and Interpretation Pipeline

In the IRL project, we have developed a three-layered pipeline to capture, enrich and allow for new interpretations of the historical data.

The aim of the first layer is to preserve the civil registers in their original form and capture the provenance of the archival record. From the digital archivist's point of view, the register pages are the main units to be preserved. The Vital Records Ontology (VRO)[2] is developed to annotate each register page and preserve the authenticity. In this layer, we converted the historical data into Linked Data and preserved them in the original order and without any interpretation.

The second layer is dedicated to creating links between the captured records and identifying the associations between them, for instance, using nominal and geographic data, individuals and familial bonds can be identified and subsequently verified by address. It also includes annotations to other standards or ontologies such as the cause of deaths. The Historical Events Ontology (HEO) was developed to enrich the registers and interlink each archival entry to constitute families.

The third layer is designed for exploring the linked records stored in the second layer from various points of interest. In this layer the data is queried which permits historians to examine the de-identified results from several perspectives. For example, the definition of maternal mortality is historically poorly defined but the pipeline permits historians to reinterpret the data in order to potentially identify additional deaths [13]. This layer permits researchers to apply different definitions, for instance the WHO's current definition of a direct maternal death is one occurring within 42 days of the delivery or termination of pregnancy [14]. Use case specific ontologies can enable the historical data to withstand multi-factorial queries for example, timeframes for deaths from puerperal sepsis (a common cause of maternal death) can be cross-referenced with the ages of the women involved to reveal patterns in maternal mortality.

4 Results and Implementation

In the IRL project, we have designed and implemented a four steps data preservation and implementation pipeline to answer historians' questions by processing the civil registration records. The proposed pipeline is implemented with linked open data

[2] http://purl.org/net/irish-record-linkage/records

standards and serveds over in the JENA Fuseki SPARQL endpoint. Figure 1 shows the four main steps of the developed pipeline. In the following section, we will describe the role of each layer in detail and present the implementation results.

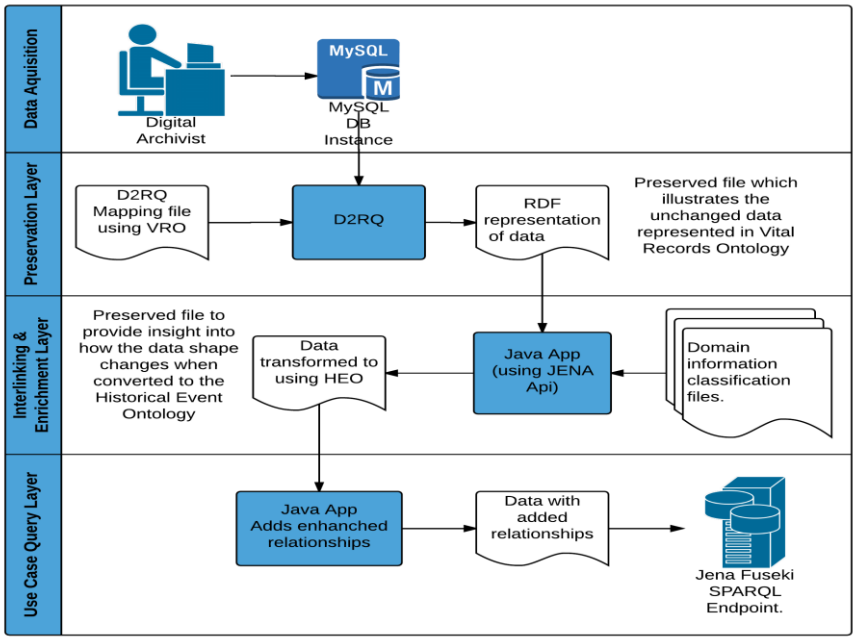

Fig. 1. Implementation of the developed semantic pipeline.

Data Acquisition

The project data consists of digitised birth, death and marriage register pages as well as a corresponding database, shared for the duration of the project by the General Register Office of Ireland. As it is presently a closed dataset, access to this data is restricted to IRL team members and no persons can be identified from the project outputs. Each register page may include up to 10 records, each one registering the birth, death or marriage of an individual. The digitised records were analysed by the digital archivists and broken down to identify all information captured in a given record and register page.

For the purpose of the project a MySQL database was created to curate a sample of the birth, death and marriage records from the Registrar Districts of Dublin South City 1 and Dublin South City 3. Digital archivists manually transcribed 444 death register page records containing 4090 death records, 15 birth register page records containing 150 birth records, and 28 marriage register page records containing 81 marriage records from 1870 and 1890. Death records have been focused on initially, as the historical research questions examine infant and maternal mortality. Using the original database for reference, as well as manual curation, relevant records were

selected and transcribed into the new database to capture all original information. The register page and the records thereon were linked to preserve the original context of record creation.

Preservation Layer

The first layer serves as a long-term digital preservation platform for digitised objects, namely Register Pages for this specific study. Register pages are transcribed verbatim in the original form and represented in Linked Data format. The aim of this layer is to provide a trustworthy platform for preserving the historical data by applying digital archival principles. Linked data structures are designed based on the provenance and archival authenticity principles.

In the preservation layer, D2RQ Mapping[3] is applied to extract the data from the MySQL database into RDF using the VRO ontology. D2RQ is a system used to treat relational databases as virtual, read-only RDF graphs. It also allows for the creation of custom dumps of the database in RDF formats for loading into an RDF store [16].

In the RDF representation, we have utilized the VRO. VRO has two basic classes for representing a digital object and its data, namely RegisterPage and Record. Births, deaths, and marriages were captured per district (within a union, within a county) as single records in a bound Register. Each register page can contain up to 10 records. The district registrar was responsible for maintaining the register and returning a true copy of all life events on a quarterly basis to the superintendent registrar for inspection and certification. The RegisterPage object encapsulates the metadata of the physical register page including dates, place, volume, stamp number as a unique identifier, district registrar, and superintendent registrar. The Record object captures data in the RegisterPage with exactly the same attributes such as name, forename, and date of birth. Because one of the projects aims is to maintain the original record by minimizing interpretation, we chose to develop a "flat" ontology, which means that most of the information that can be found on such a register page was captured as literals. A RegisterPage and a Record are linked; each record must belong to a register page and each register page can have zero or more records.Figure 2 presents linked data representation of a registry page and a death record from same page.

In the mappings, special care was taken to preserve the ability to trace information back to the source- the original records. The transcriptions included the original page numbers and unique register stamp numbers, as well as the name of the Registrar and Superintendent Registrar. As presented in figure 2, each registry page linked to the death records through 'records:withRecord' property.

[3] http://d2rq.org/ Version: D2RQ v0.8.1 - 2012-06-22

Register Page	Death Record
<http://irl.dri.ie/register_page/D4746422> a records:RegisterPage ; rdfs:label "D4746422" ; records:county "Dublin" ; records:datePageCertified "1890-01-02"^^xsd:date ; records:datePageCertifiedAsTrueCopy "1890-04-14"^^xsd:date ; records:district "South City Number 1" ; records:districtOfSuperintendantRegistrar "South Dublin" ; records:forenameOfRegistrarOnPage "752c4c2bfd8b1a4e7511e7de"records:forenameOfSuperintendantRegistrar"2042101ac741bfe43f3672e67c" ; records:pageNumber "637"^^xsd:int ; records:pageNumberOfManuscript "1"^^xsd:int ; records:quarter "1"^^xsd:int ;records:stampNumber "4746422" ; records:surnameOfRegistrarOnPage"3f6d390dbfdb5ab58b3109f6ba4" ; records:surnameOfSuperintendantRegistrar "76fecf24fdc371ebeb8e459c25d9f373" ; records:union "South Dublin" ;records:volume "2"^^xsd:int ; records:withRecord <http://irl.dri.ie/record/D4746422-67> , <http://irl.dri.ie/record/D4746422-61> , <http://irl.dri.ie/record/D4746422-62> , <http://irl.dri.ie/record/D4746422-66> , <http://irl.dri.ie/record/D4746422-68> , <http://irl.dri.ie/record/D4746422-65> , <http://irl.dri.ie/record/D4746422-70> , <http://irl.dri.ie/record/D4746422-63> , <http://irl.dri.ie/record/D4746422-69> , <http://irl.dri.ie/record/D4746422-64> ; records:yearRegistered "1889"^^xsd:int .	<http://irl.dri.ie/record/D4746422-69> a records:Certificate , records:DeathRecord , records:Record ; rdfs:label "Death of 5bd81ca81adf2879322e0ffd90b771c6db135761abfeb3b2f79fcb9ccba6 in 1889-12-29" ; records:ageLastBirthday "10 months" ; records:causeOfDeath "bronchitis" ; records:condition "bachelor" ; records:dateOfDeath "1889-12-29"^^xsd:date ; records:dateOfRegistration "1890-01-02"^^xsd:date ; records:deathCertification "Explicitly Certified" ; records:durationOfIllness "8 days" ; records:forename "5bd81ca81adf2879322e0ffd90b771" ; records:forenameOfInformant"341d3faa12b2ccbb5e772e7be" ; records:forenameOfRegistrar "d005aa409ef3b5c6bf7cb6d8b41" ; records:number "69"^^xsd:short ; records:placeOfDeath "5 Lady Lane" ; records:qualificationOfInformant "present at death" ; records:rankProfessionOrOccupation "labourer's child" ; records:residenceOfInformant "ba5ae2bd62a87e6c9af264a3ef" ; records:sex "M" ; records:surname "c6db135761abfeb3b2f79fcb9ccba6" ; records:surnameOfInformant "c4a8becd037765363048185292" ; records:surnameOfRegistrar "7cb15d9c8b537a241dca387619c" ; records:titleOfRegistrar "Registrar".

Fig. 2. An example linked data entry of Registry Page and Death Record

Interlinking and Enrichment Layer

The aim of the interlinking and enrichment layer is to facilitate the exploitation of historical data for various purposes by enabling efficient queries. Hence the data schema of the VRO were designed for preserving digitised objects as they are, it was not particularly effective querying the relations between people and events. Therefore, we developed HEO and transform another triple store for exploring the data through generations and gaining insights into longitudinal health histories.

The interlinking and enrichment layer was implemented with a Java application using JENA Api. VRO based linked records were processed and converted to the HEO based linked data. The interlinking creates associations between different pieces of information captured from register pages and allows for the reconstitution of families from the data captured in historical events. This requires the interpretation of the historical data at varying levels. The HEO is developed to represent the structure of the historical event in terms of actors, events and their relations with each other. Metadata for each level of interpretation is held with the HEO and linked to the original record to follow the provenance.

In this layer, we identified the actors of events as they are represented in the civil registration records. Depending on the historical event there are a different number of actors participating in each record. For example in a death event, four different people are identified, i.e. the person who has died, the informant, the district registrar and the superintendent registrar.

The Java app makes use of the Apache Jena API to load the data in RDF from the output of the D2RQ mapping. It works through each record, (death, birth, marriage) and develops a new data model based on the HEO. The resulting linked data contains the object classes detailed in the HEO ontology such as different types of Person (Registrar, Informant, Superintendent) and the various Event types In the next step, the actors are linked with each other according to the role they played in the historical event. Figure 3 presents the death event and the person extracted from the death record given in Figure 2.

Death Event:	**Dead Person:**
<http://irl.dri.ie/record/D4746422-69/deathEvent> a heo:DeathEvent ; heo:InformantQualification "present at death" ; heo:dateOfRegistration "1890-01-02"^^<http://www.w3.org/2001/XMLSchema#date> ; heo:deathCertification "Explicitly Certified" ; heo:eventOf <http://irl.dri.ie/record/D4746422-69/person> ; heo:placeOfDeath "5 Lady Lane" ; heo:registeredBy <http://irl.dri.ie/record/D4746422-69/registrar> .	<http://irl.dri.ie/record/D4746422-69/person> a heo:Person ; rdfs:seeAlso "http://irl.dri.ie/record/D4746422-69" ; heo:AgeAtLastBirthday "10 months" ; heo:AgeAtLastBirthdayInMins "439200" ; heo:CondAtDeath "bachelor" ; heo:dateOfDeath "1889-12-29" ; heo:forename "5bd81ca81adf2879322e0ffd90b771" ; heo:hasAtDeath <http://irl.dri.ie/record/D4746422-69/rank> ; heo:hasCauseOfDeath [heo:classifiedAs "http://purl.org/net/irish-record-linkage/historicalEvents.owl#Bronchitis" ; heo:durationOfIllness "8 days" ; heo:originalText "bronchitis"] ; heo:hasRecordFor <http://irl.dri.ie/record/D4746422-69/deathEvent> ; heo:surname "c6db135761abfeb3b2f79fcb9ccba6" .

Fig. 3. An example linked data entry of Death Even and Person record

We also enrich the original records by adding derived features. An example of these kinds of enhancements is the addition of an ageAtDeathInMinutes field. The original death records contain text in the AgeAtLastBirthday field such as "about 60 years" or "2 years and 3 months". In this form, the data would not lend itself to querying very well, for example, or would not effectively identify children who died under the age of 2. We have employed the PrettyTime:NLP library[4] to process the RDF data text fields such as heo:AgeAtLastBirthday "10 months" and convert it to minutes as heo:AgeAtLastBirthdayInMins "439200". Another type of interpretation is to enrich the existing data set with standard terminologies and ontologies. Attributes such as place name and cause of deaths can be annotated with related nomenclatures and coding systems. In this study, we examined the cause of death and

[4] http://www.ocpsoft.org/prettytime/nlp/

mapped them to different coding systems. Medical coding systems evolve over time. In 1864 all Irish Registrars were furnished with copies of a standard nosology, which identified 145 causes of death [16]. Reflecting significant advances in medical science, medical coding systems underwent a similar evolution in the period under review 1864-1913. Using the causes of death in the 1890 sample as a guide we explored the coding systems used in that time frame. To supplement the 1864 nosology we selected three available coding systems namely, the International List of Causes of Death, Revision 1 (1900) (ILCD1), the International List of Causes of Death, Revision 2 (1909) (ILCD2), and the International Classification of Causes of Sickness and Death (ICSD) [17,18,19]. The distinct cause of death is selected from the triple store, manually reviewed by the domain experts, and mapped to the available codes in ILCD1, ILCD2, and ICSD.

In HEO, we created CauseOfDeath and identified subcategories for each of them. Each subcategory is annotated with the relevant ILCD1, ILCD2 and ICSD codes. As shown in Figure 3, in the linked data repository a person object is linked with a blank node, which contains the original cause of death and duration of illness. Then individual causes of death are classified with the defined CauseOfDeath subcategories ın HEO. During this process, the HEO records are progressively enhanced to add linkages to allow for identification of individuals and to carry out normalizations such as aligning causes of death with ILCD standards. The Java app loads a custom file, which contains mappings for the domain of causes of death (as found in the data) to a standardized set of international causes of death. As it can be seen from Figure 4 heo:hasCauseOfDeath relation create a link to relevant cause of death class with heo:classifiedAs object property, and captures heo:durationOfIllness and heo:originalText as data properties of the blank node.

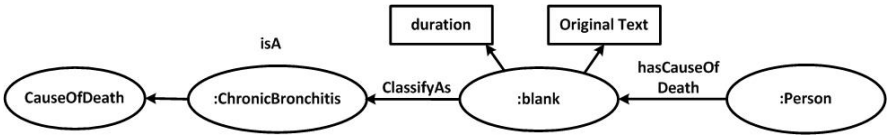

Fig. 4. Enriching the death records with ILCD1, ILCD2, and standards.

Use Case Query Layer

In the final phase, JENA Fuseki SPARQL endpoint serves to address the use cases and return the query responses. The ultimate aim of the semantic pipeline is to provide historians with tools to analyse historical events and to answer their specific research questions such as "How accurate are historic maternal mortality rates and infant mortality rates for Dublin?" Historic definitions vary for maternal and infant mortality. Infant mortality is currently defined as a death of a child before reaching the age of one, if subject to age-specific mortality rates of that period. Deaths in the first 24 hours and in the following 27 days have specific significance from the historians' perspective.

The use case query layer enables researchers to set their questions and define varying versions of concepts they are interested in. In the infant mortality use case, infant mortality is examined from multiple perspectives including the time frame of death, seasonality, location and the cause of death. Death time frame is defined with four classes; deathIn24hours, deathIn27days, infantDeath, and neoNatalDeath. Figure 5 presents SPARQL query for the deaths in 24 hours after birth. Results of queries are returned in aggregated form without disclosing any identifiable personal data. The death timeframes correspond with specific diseases and whether or not the infant was weaned too early, which can be indicative of lower socio-economic circumstances.

```
prefix heo: <http://purl.org/net/irish-record-linkage/historicalEvents.owl#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
select ?s  ?DateOfDeath ?AgeInWords
Where {
   ?s a heo:Person.
   ?s heo:AgeAtLastBirthdayInMins ?ageInMins.
   ?s heo:AgeAtLastBirthday ?AgeInWords .
   ?s heo:dateOfDeath ?DateOfDeath
   FILTER(xsd:integer(?ageInMins) <= 1440)  #1440 is 24 hours
   FILTER(xsd:integer(?ageInMins) > 0)     #Ignore records with no recorded age }
```

Fig. 5. Example use case query for deathIn24hours.

5 Discussion and Future Work

Semantic technologies and Linked Data promises many advantage for capturing, exploring and interpreting the historical data sets. The developed domain ontologies provided means for separating varying concerns, preserving authenticity and maintaining the provenance of the records. In this study, we present application principles of the semantic web technologies to create a data preservation and knowledge query pipeline by utilising the linked data together with the developed domain ontologies, namely VRO and HEO. In future we will explore more flexible and dynamic use case generation for end users and techniques to improve scalability and performance of the developed technologies.

Acknowledgements We thank the Registrar General of Ireland for permitting us to use this rich digital content contained in the vital records for the purposes of this research project. This publication has emanated from research conducted within the Irish Record Linkage, 1864-1913 project supported by the RPG2013-3; Irish Research Council Interdisciplinary Research Project Grant, and within the Science Foundation Ireland Funded Insight Research Centre (SFI/12/RC/2289). The Digital Repository of Ireland (formerly NAVR) gratefully acknowledges funding from the Irish HEA PRTLI programme.

References

1. Beyan, O., Breathnach, C., Collins, S., Debruyne, C., Decker, S., Grant, D., Grant, R., Gurrin, B. Towards Linked Vital Registration Data for Reconstituting Families and Creating Longitudinal Health Histories. In: Knowledge Representation for Health Care KR4HC 2014. Organized under the" Vienna Summer of Logic 2014" multi-conference. (2014)
2. Davies, J., Fensel, D., Van Harmelen, F. (Eds.). Towards the semantic web: ontology-driven knowledge management. John Wiley & Sons. (2003)
3. Harrower, N., Webb, S., Tang, J., Gallagher, D., Kilfeather, E., O'Tuairisg, S., Collins, S. Developing the Irish National Trusted Digital Repository for the Humanities and Social Sciences: an interdisciplinary approach. OR2013. (2013)
4. Hunter, J., Choudhury, S. PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. International Journal on Digital Libraries 6.2: 174-183. (2006)
5. Lee, K.H.; Slattery, O.; Tang, X.; Lu, R.; McCrary, V. The state of the art and practice in digital preservation. Journal of Research-National Institute of Standards and Technology 107.1: 93-106. (2002)
6. Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., Lorés, J. Semantic web technologies for a digital newspaper archive. The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 445-458. (2004)
7. Breathnach, C., O'Halpin, E. Registered 'unknown' infant fatalities in Ireland, 1916–32: gender and power. Irish Historical Studies, 38(149), 70-88. (2012)
8. Breathnach, C., O'Halpin, E. Scripting blame: Irish coroners' courts and unnamed infant dead, 1916–32. Social History, 39:2, 210-228. (2014)
9. Sheth, A.P., Ramakrishnan C. Semantic (Web) technology in action: Ontology driven information systems for search, integration, and analysis. IEEE Data Engineering Bulletin 26.4: 40. (2003)
10. Bizer, C., Heath, T., Berners-Lee, T. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts: 205-227. (2009)
11. Meersman, R., Debruyne C. Hybrid ontologies and social semantics. Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on. (2010)
12. Debruyne, C., Meersman, R. Semantic interoperation of information systems by evolving ontologies through formalized social processes. Advances in Databases and Information Systems. Springer Berlin Heidelberg. (2011)
13. Kippen, R. Counting nineteenth-century maternal deaths: the case of Tasmania. Historical Methods: J. of Quantitative and Interdisciplinary History, 38:1, 14-25. (2005)
14. WHO. International Classification of Diseases and Related Health Problems (Geneva: World Health Organization) (1992)
15. Great Britain. Lord Lieutenant and Privy Council of Ireland. Registrar General of Ireland Registration of deaths in Ireland: a statistical nosology, comprising the causes of death, classified and alphabetically arranged with notes and observations, Dublin. (1864)
16. Bizer, C., Cyganiak, R. D2r server-publishing relational databases on the semantic web. Poster at the 5th International Semantic Web Conference. (2006)
17. Int. List of Causes of Death, Rev.1 (1900). http://www.wolfbane.com/icd/icd1h.htm
18. Int. List of Causes of Death, Rev.2 (1909) http://www.wolfbane.com/icd/icd2h.htm
19. Department of Commerce and Labor, Bureau of Census. International Classification of Causes of Sickness and Death. Washington Government of Printing Office. (1910)